

Combinatorial Sample Pooling Assigns Rare DNA Variants to Known Individuals

April 15, 2013

ME Arnegard

The vast majority of traits - both adaptive phenotypes and disease states - are genetically complex, being influenced by many genes. A classic example, recognized long ago, is human height variation. In the last ten years, genome-wide association studies (GWAS) have revealed complex genetic architectures for several medically important traits, such as duration of the reproductive lifespan (*e.g.*, Chen *et al.*, 2012).

In the GWAS approach, scans are conducted across the genome for statistically significant relationships between a phenotypic trait of interest and 'common' genetic variants (*i.e.*, alleles with frequencies of at least 5%) at millions of different genetic markers, typically single nucleotide polymorphisms (SNPs). Even among the most intensively studied traits, however, the suite of contributing markers accounts for only a fraction of the genetic heritability known for a given trait. The remaining, 'missing heritability' cannot be detected by current methods, though geneticists are just as sure of its presence as cosmologists are of the dark matter that permeates and structures the universe (Manolio *et al.*, 2009). 'Rare' genetic variants (<5% frequency) which have important effects on complex traits could solve part of the problem of 'missing heritability', though discovering these rare alleles poses technical and financial challenges.

In a recent paper published by the laboratory of Dr. Christopher Carlson (Public Health Sciences Division), Dr. Christina Chen and colleagues describe a rare-variant detection algorithm they developed, called *V-Sieve*. This algorithm is based on an ingenious strategy of pooling DNA across multiple sampling dimensions (see figure). Chen *et al.* demonstrate the utility of their new approach by re-sequencing the C-reactive protein (CRP) locus of individuals from the 'Coronary Artery Risk Development in Young Adults' (CARDIA) cohort. This cohort was established in 1984 to prospectively study biological factors in young adults that influence the development of coronary artery disease up to 20 years later. The authors were particularly interested in CRP, which plays a role in inflammatory responses and the killing of necrotic cells, because serum CRP levels are now taken as valuable predictors of acute coronary events (Pepys *et al.*, 2006).

Despite rapid advancements in DNA sequencing technology, individually sequencing entire genomes of many study participants still remains prohibitively costly when it comes to detecting rare

genetic variants underlying complex traits. Instead, Chen *et al.* amplified the 6 kilobase *CRP* locus in 2,283 individuals from the CARDIA cohort using long range PCR. The investigators created multiple pooled samples, containing amplified DNA from 96 individuals each, to help them detect rare DNA variants using deep sequencing at 700 - 900 fold coverage. Dr. Carlson's group developed a metric (called *STAT*) to statistically distinguish genuine rare polymorphisms from background technical noise across pools, and they used a combinatorial pooling scheme to simultaneously assign these rare variants to known individuals within pools. The authors then validated these aims by simulating null distributions of *STAT* and by conducting targeted re-sequencing in study participants identified as rare-*CRP*-allele carriers.

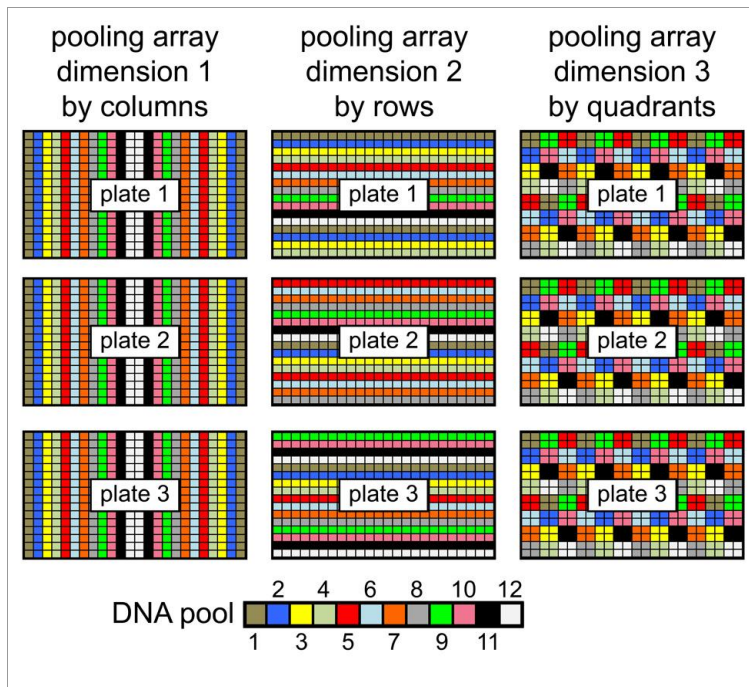
Using *V-Sieve*, the authors identified 84 candidate polymorphisms that had not been previously documented in the *CRP* region. Most of these new genetic variants, including two tri-allelic loci, were confirmed by the authors' targeted re-sequencing efforts (positive predictive rate = 93%). The majority of the new *CRP* variants occurred at frequencies of <1%, suggesting that rare polymorphisms in and around *CRP* might predominate over polymorphisms with a minor allele frequency (MAF) of at least 5%. In fact, the *V-Sieve* algorithm was successful in identifying new polymorphisms with a MAF of just 0.02%. Some of these polymorphisms could enhance the predictive value of CRP screening for healthy individuals and cardiology patients alike. In addition, the authors' findings suggest that ultra-rare alleles account for part of the missing heritability of complex traits, though other potential factors accounting for this 'dark matter' of genetics include epigenetic effects and the influences of high-order genetic interactions on trait heritability (Manolio *et al.*, 2009).

[Chen CT, McDavid AN, Kahsai OJ, Zebari AS, Carlson CS](#). 2013. Efficient identification of rare variants in large populations: deep re-sequencing the *CRP* locus in the CARDIA study. *Nucl. Acids Res.*, Epub ahead of print, 13-Feb-2013, doi:10.1093/nar/gkt092.

Also see: [Chen CT, Fernández-Rhodes L, Brzyski RG, Carlson CS, Chen Z, Heiss G, North KE, Woods NF, Rajkovic A, Kooperberg C, Franceschini N](#). 2012. Replication of loci influencing ages at menarche and menopause in Hispanic women: the Women's Health Initiative SHARe Study. *Hum. Mol. Genet.* 21:1419-32.

[Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM](#). 2009. Finding the missing heritability of complex diseases. *Nature* 461:747-53.

[Pepys MB, Hirschfield GM, Tennent GA, Gallimore JR, Kahan MC, Bellotti V, Hawkins PN, Myers RM, Smith MD, Polara A, Cobb AJ, Ley SV, Aquilina JA, Robinson CV, Sharif I, Gray GA, Sabin CA, Jenvey MC, Kolstoe SE, Thompson D, Wood SP](#). 2006. Targeting C-reactive protein for the treatment of cardiovascular disease. *Nature* 440:1217-21.



Adapted from Fig. 1 of Chen et al.

Combinatorial pooling scheme of Chen et al. Each plate contained DNA from 384 different CARDIA cohort participants. The DNA was pooled in three different ways across plates, resulting in three different sets of pools containing 96 individuals each. In the actual study, there were 24 pools, rather than just the 12 shown here. Pooled PCR products were then used to generate a barcoded library, which was subjected to NexGen sequencing. Thus, each participant was sequenced once in each pooling dimension and three times across all 72 pools (24 x 3). A single individual carrying an ultra-rare (i.e., singleton) allele in the population can be 'mapped' in pooling space by the unique pattern of wells, across pooling dimensions, in which the singleton appears. Although the identification of individuals carrying rare alleles present in two or more people is often possible with just three dimensions, unambiguous identification of these slightly more common alleles is attainable by incorporating additional pooling dimensions.